“一带一路”倡议下的Twitter文本主题挖掘和情感分析*

赵常煜[1] 吴亚平[2] 王继民[1]
[1] 北京大学信息管理系 北京 100871 [2] 北京大学图书馆 北京 100871

摘要：［目的/意义］“一带一路”倡议的提出引起了国内外广泛的关注，众多国家的用户在最具代表性的社交媒体Twitter中表达观点、发表评论、相互讨论。从推文中挖掘得出世界对“一带一路”的讨论主题和情感倾向，有助于为政府机构优化宣传策略，增加“一带一路”倡议的曝光度、关注度提供参考。［方法/过程］采集2017年与“一带一路”相关的6万余条推文，分别按照中文和英文进行数据预处理、数据描述、主题挖掘、情感分析，并实现主题和情感的交叉分析，得出结论。［结果/结论］2017年的推文主题主要围绕5月份的“一带一路”高峰论坛。其中，中文推文更关注高峰论坛的筹划和实施，以及安全问题、领导层的访问等方面的内容，情感值的波动较大，特别是安全问题上的消极情绪波动很大。英文推文则更关注举办高峰论坛的事实以及论坛所带来的经济效应，情感波动较小，经济方面的情感值是积极占比明显高于消极和中立的情感值。

关键词：“一带一路” Twitter 主题挖掘 情感分析
分类号：TP391.1
DOI：10.13266/j.issn.0252-3116.2019.19.012

2013年9月和10月，中国国家主席习近平先后提出共建“丝绸之路经济带”和“21世纪海上丝绸之路”（以下简称“一带一路”）倡议，受到了国际上的高度关注，得到了有关国家的积极响应。社交媒体是人们获取信息的重要来源和表达观点、相互交流的重要窗口。至今，社交媒体不再仅聚焦于人的生活、娱乐，在政策观点的宣传方面也体现出越来越重要的作用，逐渐承担参政工具、商业广告平台、讨论社区等角色。美国总统特朗普常在国外社交媒体中宣传自己的政策、发表与政治活动相关意见，有些话题引起了广泛的社会舆论。Twitter作为一种微博客型社交媒体，至2017年底，已支持中文、英文等全球34种语言，累积激活用户数量达到3.6亿，包括政治、体育、娱乐等多个领域的领头人物。众多国家用户在Twitter上对“一带一路”倡议展开了热烈的讨论，2017年全年共有10万余条相关推文。在这一背景下，政府机构如何合理利用社交媒体，加大国家倡议的宣传力度，增加曝光度，激发更多的讨论十分必要。基于此，本文运用多种方法从推文评论数据挖掘人们对“一带一路”倡议的讨论主题和情感倾向，得出交叉性结论，以丰富相关研究，为“一带一路”倡议的宣传报道提供参考。

1 相关研究

1.1 围绕“一带一路”的研究现状

截至2018年4月，中国知网共有4万多篇“一带一路”相关文章，涉及政策分析、经济效应分析、科研合作分析等多方面。其中，文本分析相关的文章于2015年首次发表，数量较少，多数以新闻媒体报道为研究对象，涉及《中国日报·非洲版》《华盛顿邮报》等，挖掘得出语义结构特征、主题观点。如黄炎秋基于《中国日报·非洲版》文本数据，讨论了公共外交与传播新常态矛盾等问题，分析过程中发现了“一带一路”议题摆脱了“一边倒”的现象，成为了国际性话题，最后为宣传报道提出了策略和建议[1]。朱桂生等学者利用了美国的《华盛顿邮报》文本数据，对“一带一路”主题下相关报道进行了批评性话语分析，从文本、话语实践和社会实践3个层面入手，揭示了美国媒体把中国的“一带一路”倡议塑造成了一种殖民扩张、重利轻义的霸权形象[2]。着眼于社交媒体文本的相对较少，国内数据源主要基于新浪微博开展，如采集2013－2016年约36

* 本文系国家社会科学基金项目“‘一带一路’沿线国家互联互通水平综合评价研究”（项目编号：16BTQ057）研究成果之一。
作者简介：赵常煜（ORCID：0000－0001－6780－1070），硕士研究生；吴亚平（ORCID：0000－0002－4242－2434），馆员；王继民（ORCID：0000－0002－3573－7788），教授，博士生导师，通讯作者，E-mail：wjm@pku.edu.cn。
收稿日期：2018－12－11 修回日期：2019－03－21 本文起止页码：119－127 本文责任编辑：刘远颖

图书情报工作
第63卷 第19期 2019年10月

万条新浪微博文本数据,利用空间自相关分析方法,证明核心城市和边缘城市之间的倡议响应差异,提出了优化宣传空间结构等建议[3]。国外数据源主要基于Twitter开展,如采集2015年"一带一路"主题下2 000余条推文,进行地域划分和关键词统计,分析信息传递过程[4]。然而目前的研究局限于结构化的传统新闻文本,存在数据量较少、数据来源比较单一、分析维度较少等不足。

1.2 主题挖掘方法研究进展

主题挖掘是利用文本集合中文本特征项之间的关联关系发现研究主题的过程。分析主题在时间维度上的演化分析,可以明确主题发展脉络,寻找创新点。传统的主题挖掘方法主要有词频分析法、共词分析法和引文分析法等。词频分析法是基于齐普夫定律,通过关键词或主题词的出现频次来确定主题的方法。虽简单易用,但高频词和低频词具有非常强的主观性,会导致主题范围比较广,主题难以归一等问题。共词分析法是基于统计思想,查看两个关键词在同一篇文献中的共现关系,兼顾了词频和词与词之间的关系,但低频词不易被纳入到主题的讨论之中。引文分析法是基于引用和被引用的关系,通过引用率、引用耦合和同被引等指标进行主题划分,但面临着引用关系复杂、引用格式不统一等问题。整体来看,传统的主题演化分析方法虽易操作,应用比较广,但主观性较强,研究结论较浅。

之后出现了结合机器学习和自然语言处理的复杂模型,LSI、PLSI[5]、LDA等都属于这个范畴。隐含狄利克雷分布模型(Latent Dirichlet Allocation, LDA)是D. M. Blei等[6]在2003年提出的确定一组文档的共同主题的技术,认为一篇文章的每个词都是通过"以一定概率选择了某个主题,并从这个主题中以一定概率选择某个词"的方式得到主题分类[7]。2010年前后有学者将LDA模型应用于社交媒体上,M. Michelson等[8]利用LDA模型研究Twitter用户所关注的主题内容,Y. S. Hwang等[9]利用LDA主题模型研究了意见领袖讨论主题的规律和方法。Y. Hu等[10]应用LDA模型分析时事新闻的社交媒体评论数据,得出用户观点。随后又出现了基于PLSA和LDA的改进模型。如Q. Mei等[11]对PLSA模型进行了改良,将词语的上下文的信息应用到PLSA模型上,称为CPLSA(Contextual Probabilistic Latent Semantic Analysis),S. Moghaddam等[12]提出的增量潜在狄利克雷分类(ILDA),在LDA模型的基础上,添加文本特征参数,提高了主题聚类的准确性,主要应用于从评论中抽取主题及得分。此外,还出现了动态主题模型(DTM)[13]和在线潜在狄利克雷分类(OLDA)[14]等。可见众多学者对LDA模型进行了广泛、深入的研究,模型得到了较好的发展,因此,本文选取LDA的方法进行推文的主题挖掘。

1.3 情感分析方研究进展

情感分析是从文本数据中识别出用户主观的情感、观点和态度的过程[15]。在舆情监控和信息预测等方面应用较广。最初的社交媒体情感分析就是基于Twitter的社交媒体数据开展的[16-18]。J. Bollen等[19]基于Twitter数据把情感分成6个情感维度,分析出了每天最具代表性的情感;P. S. Dodds等[20]从情感分析的角度尝试解释了人们感到幸福的规律。情感分析方法主要可分为基于情感词典的情感分析方法和基于机器学习的情感分析方法。基于情感词典的情感分析是从待测文本中提取特征词后,在情感词典中查找该特征词的情感值,根据累加的情感值进行情感分类的方法[21]。在情感词典的选择上,一般有两种方式:一种试引用已有的情感词典,如HowNet词典[22]、SentiWordNet、Inquirers等[23];另一种是通过研究数据自行构建词典,如R. Feldman等[24-25]学者在已有的情感词典的基础上,利用部分人工标注和Bootstrapping的方式提取情感词。基于机器学习的情感分析方法则先基于文本集训练得到分类器,再基于分类器实现对新文本的分类[26]。随着人工智能和深度学习的发展,不少学者把深度学习的技术运用到情感分析中。B. Pang等[27]首次利用机器学习的方法对电影的评论文本做了情感分析。张志华[28]在情感词向量的基础上利用卷积神经网络模型进行情感分析,通过英文文本做了实证研究,结果显示基于深度学习模型的分类结果占一定的优势。

然而对于非结构化的文档,如微型博客、社交媒体等,基于机器学习方法的分类效果并不理想。由于推文限制在140个字以内,长度较短,通常表达1-2个句子,其中包含表达情感的情感符号和网络用语,机器学习的方法对这些符号和用语相对不敏感。基于机器学习的情感分析还需要依赖大量语料的训练和人工干预,耗时间比较长,因此本文对推文的处理选用基于情感词典的情感分析方法。

2 推文主题挖掘和情感分析方法

2.1 主题挖掘方法

文章基于LDA主题模型算法挖掘主题。目前普

遍认为应用 LDA 的最关键的在于最佳主题数目的确定,LDA 主题抽取的效果和潜在主题数目有直接关系[29]。国内外学者提出了最小困惑度算法、HDP 算法、贝叶斯算法[30]等多种确定最优主题数量的方法。综合考虑,本文采用由 R. Michael 等 2015 年提出来的 Coherence 方法作为评价模型好坏的评价标准,通过选取 Coherence 最大的模型来确定主题的最佳数目。主题挖掘的流程如图 1 所示,经过数据预处理、分词后,建立词频特征构建文档－单词矩阵,利用 Coherence 值确定最佳的主题数目,构建 LDA 模型,挖掘得出主题。

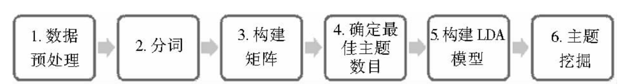


图 1　主题挖掘流程

2.2　情感分析方法

应用基于情感词典的情感分析方法时,本文在现有的研究基础上,构建和扩充了情感词典。针对中文推文和英文推文,分别采用了如下方法:

(1)中文推文的情感分析基于大连理工大学情感词汇本体库对情感词进行扩充。大连理工情感词汇本体库将情感分为“乐、好、怒、哀、惧、恶、惊”7 类,并定义了情感的强度,但未涉及情感词在句子中与程度副词、否定词、表情符号等之间的关系。然而在中文句式中,不仅否定词和程度副词在情感词的前后位置关系会影响情感强度,而且否定词的出现次数也会影响整体情感值。笔者综合考虑句子中否定词、程度副词对情感词的作用,分别构建否定词词典和程度副词词典,借鉴杨希[31]的 6 种情感词组合方法(见表 1),综合考虑否定词、程度副词之间的相互作用,计算推文的情感值。

表 1　情感词组合模式

序号	情感词句式
1	[情感词]
2	[否定词] + [情感词]
3	[程度副词] + [情感词]
4	[否定词] + [程度副词] + [情感词]
5	[程度副词] + [否定词] + [情感词]
6	[否定词] + [否定词] + [情感词]

在计算每条推文情感值时,以每个情感词为基准,发现否定词和程度副词的位置关系,累加文本中的情感值,7 个维度情感计算公式如下:

$$E_{TW-i} = \sum_{j=0}^{K} e_i \cdot (-1)^N \cdot P \qquad \text{公式(1)}$$

其中,i 表示七大情感类别中的某一类,E_{TW-i} 表示一条推文在 i 类的情感值,K 表示一条推文中出现的所有情感词个数,ei 表示一个情感词在 i 类上的情感强度,N 表示与该情感词相关的否定词个数,P 表示程度副词的加权值。

(2)英文推文的情感分析基于 Wordnet 构建的 SentiwordNet3.0 对情感词典进行扩充。SentiwordNet3.0 目前包含 117 659 个词。利用随机漫步模型,为每个 Synset 下的词赋予了 PosScore(正向情感值)和 NegScore(负向情感值)。英文情感分析同样构建否定词和程度副词词典,并使用在中文情感分析过程中构建的表情符号词典。每条英文推文的情感值计算公式如下:

$$E = \sum_{j=0}^{K} e_{pn} \cdot (-1)^N \cdot P + Pos\text{-}Neg \qquad \text{公式(2)}$$

其中,E 表示一条英文 Twitter 的正向或负向情感值,K 表示一条推文中出现的所有情感词个数,e_{pn} 表示一个情感词在正向和负向的情感强度,N 表示与该情感词相关的否定词个数,P 表示程度副词的加权值,Pos 表示积极表情符号个数,Neg 表示消极表情符号个数。

(3)为了更明确地呈现推文的情感倾向,计算每条推文三元情感极性。三元情感极性即积极、中立、消极,具体计算公式如下:

$$E_p = \sum_{j=0}^{K} e_p \cdot (-1)^N + Pos\text{-}Neg \qquad \text{公式(3)}$$

其中,E_p 表示一条推文的情感极性,1 表示积极、0 表示中立、-1 表示消极。在 i 类的情感值,e_p 表示一个情感词的极性,N 表示与该情感词相关的否定词个数,Pos 表示积极表情符号个数,Neg 表示消极表情符号个数。

2.3　主题—情感交叉分析

主题—情感分析是结合主题挖掘和情感分析的结果,得到不同主题下的情感倾向随时间变化的趋势。具体实现过程为:通过主题挖掘得出每一条推文的主题概率分布,通过情感分析得出每条推文在不同主题下的情感值,按照主题进行累加计算,最终得出随时间不同主题下的情感值变化。

3　实证研究

3.1　推文采集与数据概况

以“One Belt One Road”“OBOR”和“一带一路”3 个关键词为限定词,共采集到 2017 年 1 月 1 日－12 月 31 日间 102 029 条相关推文,数据样例如表 2 所示:

表 2 推文数据爬取结果(样例)

账号	发布时间	推文	回复数	转载数	点赞数
@ water_ futures	31 - Dec - 17	#China's One Belt One Road projects in …	0	1	0

为保证实验结果的准确性,删除由于 Twitter 模糊检索带来的低相关性数据及法文、西班牙文等其他语种数据,最终保留了 63 907 条推文进行深入分析,其中中文文本为 11 457 条,英文文本为 52 450 条,由 23 706个 Twitter 用户发出。这些用户来自于不同的地区,其身份也不同。已知的地域信息中,美国和加拿大的北美地区用户占 18% ,印度尼西亚雅加达占 8% ,马来西亚吉隆坡占 4% ,中国北京占 4% 。从采集到的用户数据来看,用户可以分为媒体、记者、政治家、专家、拥护者、一般个人用户等类型,并表现出如下特征:媒体型用户在抛出话题,记者型和政治型、专家型、一般个人型用户共同推动话题的讨论,拥护者一向发表支持或积极的观点。从发文数量来看,媒体型的账号发文量大于个人账号。2017 年"一带一路"主题下月度推文数量分布情况如图 2 所示:

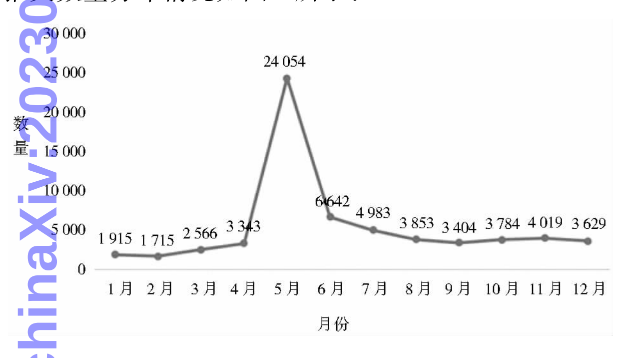


图 2 "一带一路"主题下月度推文数量分布

可以看出,2017 年 1 - 4 月"一带一路"主题推文数量稳中有升; 5 月中国举办了"一带一路"国际合作高峰论坛,相关话题量大幅上升,达到峰值;6 月推文数量锐减,随后较为平稳,但整体水平较 1 - 4 月份有所上升,一定程度上反映了 5 月份"一带一路"国际合作高峰论坛的举办,引起了 Twitter 用户对"一带一路"的高度、集中的关注,带动了后几个月对"一带一路"话题的讨论。

为了提高中文推文的分词精度,收集与"一带一路"相关的 360 篇官方新闻报道,利用"新词发现"和 TF-IDF 算法,提取每篇报道的主要关键词,得到自定义分词词典,共包含 8 242 个词。利用结巴分词工具进行分词。对于英文推文,采用词形还原的方法,考虑一个词在文章中的词性(Part-of-speech)再对单词进行还原。完成分词后,提取中、英文的推文中的名词性词语,根据词频得到推文词云,如图 3 所示:



图 3 "一带一路"主题下中、英文推文词云

可以看出,中文的词频明显大于英文的词频,高频词语多数与"一带一路"高峰论坛相关,排名前十的中、英文高频词如表 3 所示:

表 3 中、英文推文词频统计

排名	中文		英文	
	词汇	词频	词汇	词频
1	一带一路	11 426	onebeltoneroad initiative	984
2	中国	3 499	new silk road	403
3	合作	1 672	onebeltoneroad project	309
4	国际	1 551	onebeltoneroad summit	236
5	国家	1 282	pakistan	227
6	高峰论坛	1 181	china	314
7	北京	1 096	central asia	208
8	建设	834	silk road	193
9	经济	727	economic corridor	182
10	发展	708	sri lanka	156

3.3 主题挖掘结果

通过实验分别计算中文推文 3 - 15 个主题下的 Coherence 值,找出该值最大的主题数,作为 LDA 最佳主题数目。实验表明,当主题数为 6 时,Coherence 值最大,因此中文推文主题数定为 6。基于 LDA 模型分析得出"一带一路"与朝鲜半岛和平问题、"一带一路"与经济问题、"一带一路"与高峰论坛之高层访问、"一带一路"与高峰论坛之合作与项目、"一带一路"与外交战略、"一带一路"与国内热点,共六大中文推文热议主题,其相关的关键词分别见表 4。

"一带一路"与朝鲜半岛和平问题主题,凸显了 2017 年 4 月 29 日朝鲜试射导弹带来的紧张的周边局势,周边国家如韩国、俄罗斯十分关注,也说明了"一带一路" 的发展一定程度上依赖于周边国际环境。"一

表 4　2017 年中文推文主题挖掘结果

主题	关键词
“一带一路”与朝鲜半岛和平问题	一带一路、国际、朝鲜、导弹、韩国、团长、文明、安保、局势、社会、发展、全球、和平、国资委、朝鲜半岛、建设、能量、金正恩、工作、代表、俄罗斯、生态、作者、时事等
“一带一路”与经济问题	一带一路、中国、建设、会见、金融、国家、李克强、峰会、总理、股市、合作、发展、老百姓、宣传、经济、人民币、资金、美国、货币、俄罗斯、领导人等
“一带一路”与高峰论坛之高层访问	一带一路、中国、国际、高峰论坛、国家、北京、美国、倡议、主席、发展、论坛、总统、会议、印度、日本、代表、领导人、普京、峰会、协议等
“一带一路”与高峰论坛之合作与项目	一带一路、经济、峰会、建设、国家、发展、合作、全球化、圆桌、市场、投资、政治、战略、菲律宾、新西兰、欧洲、智库、巴基斯坦、教育、文化等
“一带一路”与外交战略	一带一路、中国、宪章、国家、投资、政权、大陆、合作、挑战、声明、政策、铁路、基础设施、会见、外交部、世界、贸易、机遇、国际、战略、项目、建设等
“一带一路”与国内热点	一带一路、北京、代表团、中国、国家、湖北、广西、李克强、政府、浙江、卫星、电商、建设、山东、全国、新闻等

带一路”与经济主题，凸显了经济发展、金融合作的重要工作，“一带一路”为人民币的国际化带来了机遇，资金融通与老百姓生活的方方面面息息相关。第三个和第四个主题分别专注于“一带一路”高峰论坛的高层访问和合作与项目方面，凸显了国际的关注与多个国家元首的参与以及峰会推动的政治、投资、教育、文化等多方面的全球化合作。“一带一路”与外交战略主题，凸显了基础设施、铁路、贸易、投资等重要外交领域，设施联通也是“一带一路”建设的优先领域，合作项目同时面临机遇和挑战。“一带一路”与国内热点，凸显了北京等国内省、市的积极响应，为更多的国内电商“走出去”创造了机遇。

计算英文推文随主题数变化的 Coherence 值，实验得出当主题数为 4 的时候，Coherence 值取得最大为 0.79，因此限定主题数为 4 进行 LDA 主题挖掘，基于 LDA 模型分析得出“一带一路”与 CPEC、“一带一路”与对外合作、“一带一路”与经济效应、“一带一路”与对外政策，共四大英文推文热议主题，其相关的关键词分别如表 5 所示：

表 5　2017 年中文推文主题挖掘结果

主题	关键词
“一带一路”与 CPEC	china cpec、global economy、pakistan、onebeltoneroad project、grand strategy、regional connectivity、official cpec、beltandroad summit、long-erm strategy、china relation、infra project 等
“一带一路”与对外合作	infrastructure project、india、japan、44bil project、sino indian relation、chinese firm、economic growth plan、russia、international airport、foreign policy plan 等
“一带一路”与经济效应	onebeltoneroad initiative、eastern europe、global power、economic development、model project、joint connectivity project、global trade、financial express、economic cooperation 等
“一带一路”与对外政策	road initiative、new silk road、china pakistan、central asia、foreign policy、economic corridor、primary focus、foreign affair、chinese investment 等

“一带一路”与 CPEC 主题下，CPEC 是中国—巴基斯坦经济走廊(China-Pakistan Economic Corridor)的缩写，被称为贯穿南北丝绸之路的枢纽，以加强中巴之间交通、能源、海洋等领域的交流与合作，促进共同发展。中国石油天然气公司是中巴天然气管道的建设的重要力量，为区域联通和更长远的战略合作做了良好的示范。“一带一路”与对外合作主题，再次凸显了基础设施合作的重要性，俄罗斯、印度等合作国家及国际关系受到了更多的讨论。“一带一路”与经济效应主题，凸显了全球电力、全球贸易、经济合作、联合联通等“一带一路”全球化发展的经济蓝图。“一带一路”与对外政策主题，则涉及“新丝绸之路”“经济走廊”等整体性外交倡议以及区域间合作政策等，受到了广泛关注与热议。

3.4　情感分析结果

通过情感极性判断文本的肯定、否定、中立三元情感态度，量化打分得出情感强度，2017 年“一带一路”主题下每月中文推文三元情感分析结果如图 4 所示：

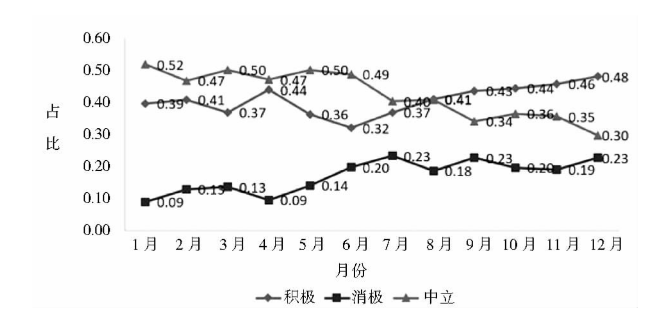


图 4　2017 年“一带一路”主题下每月中文推文三元情感分析结果

图书情报工作
第63卷 第19期　2019年10月

从趋势上来看，中文用户对“一带一路”的态度中积极和中立均高于消极。1－7月中立居高，8月份为转折点，从9月份后用户中立的情感逐渐分明，积极情感稳步、直线上升，在12月份达到最高，为0.48。消极情感也有少量增加且趋于稳定。

英文推文的情感分析结果如图5所示：

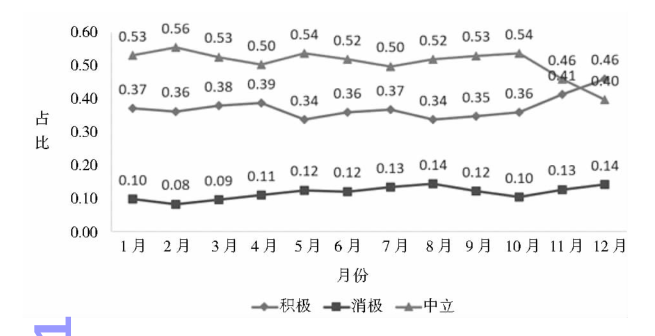


图5　2017年“一带一路”主题下每月英文推文三元情感分析结果

相比来看，英文推文中表现的情绪比较稳定，全年中立、积极、消极的情感值平均值分别为0.51、0.37、0.12。5月份举行高峰论坛前后的情感变化不大，10月份十九大的召开得到了外媒的关注，高峰论坛的举办凸显了“一带一路”的影响力，其发展再上新台阶，正是在10月份，情感值有明显的转折点，情感倾向更加明确，积极情感一直稳中有升，在12月超越了中立的情感占比，消极情感也有相对较少的小幅上升。

3.5　主题情感交叉与演化分析

随着时间的推移，用户讨论话题在不断变化，不同主题下的情感值也在变化。2017年中文推文主题的情感演化结果见图6。

可以看出，“‘一带一路’与朝鲜半岛和平问题”主题的情感波动比较大，4－6月和8－9月，消极的情感值明显上升，这一变化的原因推测可能与以下几点有关：2017年9月份朝鲜第六次核试验，受到了来自全世

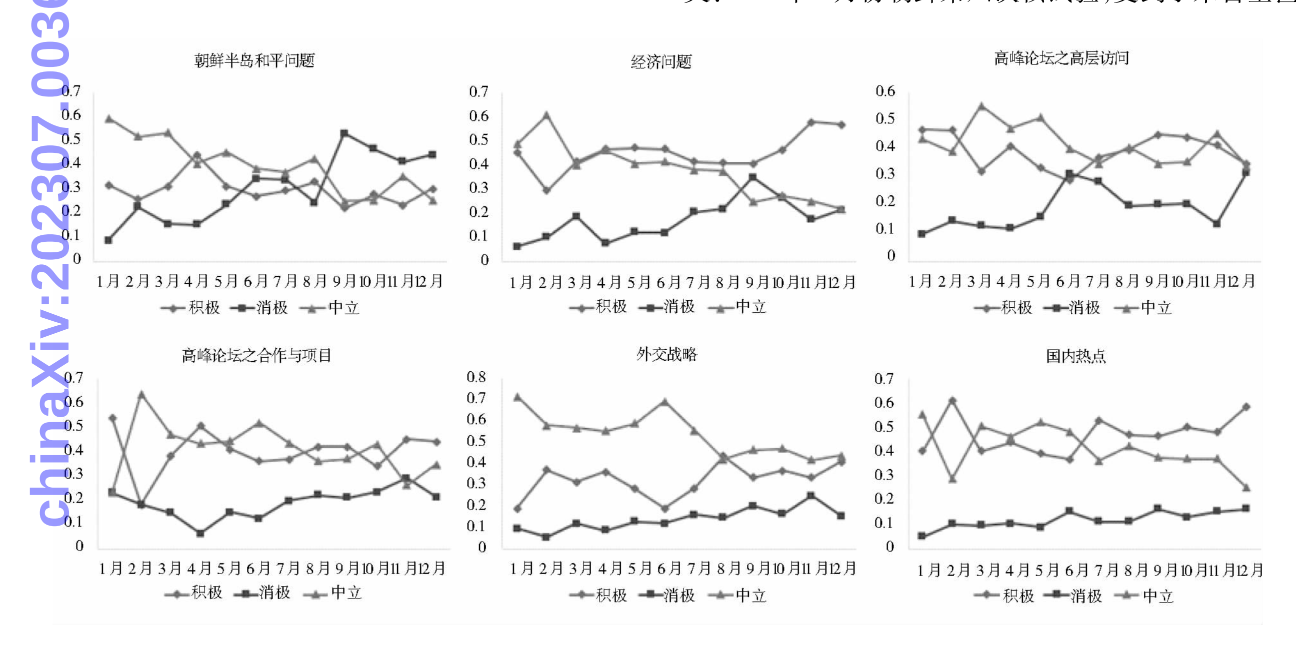


图6　中文推文主题在不同月份（横轴）的情感类型占比（纵轴）变化

界各国的谴责，中国政府也对朝鲜表示遗憾和反对；朝鲜导弹问题，除了9月份的核试验之外，2、3、4、9、11月朝鲜发射了导弹导致了朝鲜半岛周围的紧张局势。“‘一带一路’与经济问题”主题的积极情感值从3月份到年底一直高于中立和消极，到了年底积极情感值的占比超过50%。“‘一带一路’与高峰论坛之高层访问”和“‘一带一路’与高峰论坛之合作与项目”两个主题都与高峰论坛相关，高层访问相关主题在6月份出现了消极的峰值，6月份日本首相安培表示日本加入“一带一路”大计划当中，不少中文推文用户对此表示了一定的反对。合作方面，持积极和中立态度的用户占据多数，而到了年底消极态度的推文数量占比有较大的提升。2017年11月份，国外媒体开始担忧斯里兰卡、巴基斯坦等国家，他们认为“一带一路”上的中国资本过多引进到本国，过度经济依赖导致失去本国决定权。“‘一带一路’与外交战略”主题的情感值中，中立态度持主流，最多的时候达到了70%以上，可能是多数用户宣传了中国与多数国家签署了协议，也进行了外交战略方面事实的转载，较为中立。“‘一带一路’与国内热点”主题下包括了很多地名和地区信息，该主题下推文的内容多数为国内某个地区在“一带一路”项目中取得的成果和宣传。这一主题下的积极情感

赵常煜，吴亚平，王继民．“一带一路”倡议下的 Twitter 文本主题挖掘和情感分析[J]．图书情报工作，2019，63（19）：119－127．

值在 5 月份后占据了主流，消极的情感占比较低且波动较小，中文用户对“一带一路”与国内发展较为看好。

2017 年英文推文的主题及其情感演化结果如图 7 所示：

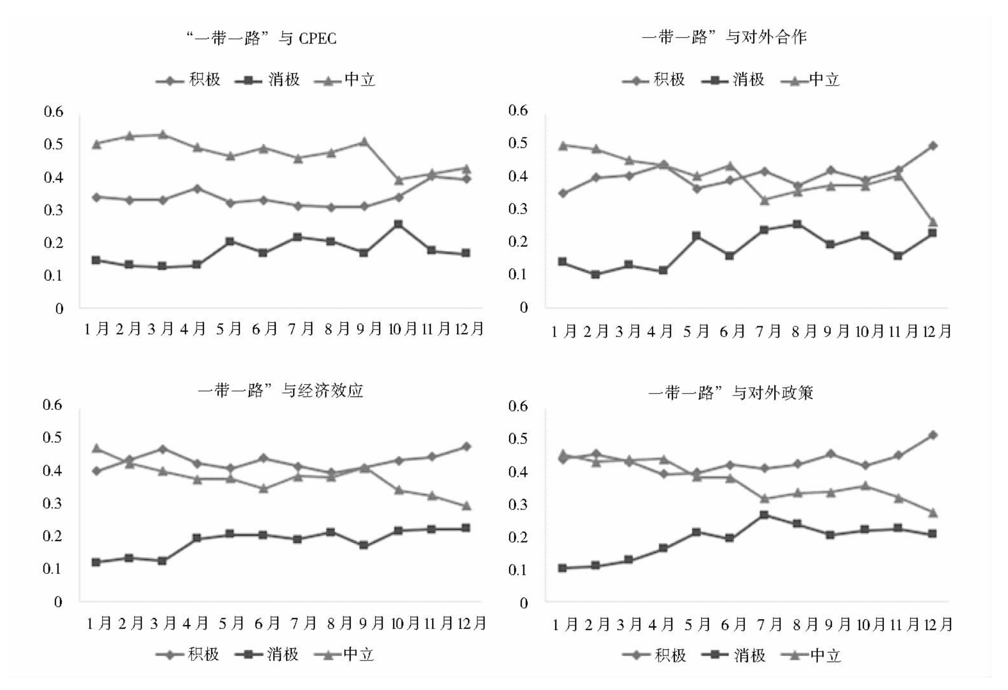


图 7　英文推文主题在不同月份（横轴）的情感类型占比（纵轴）变化

可以看出，“‘一带一路’与 CPEC”主题的中巴合作，引起了国外英文推文用户的兴趣，多数关注中巴经济走廊的经济效果，国内外对 CPEC 的报道比较偏乐观积极，如法国媒体[32]在 7 月份估算 CPEC 在当地创造了 30 万个岗位等信息。该主题下的情感波动较小，中立占多数，10 月份后积极情感值有整体性提升。“‘一带一路’与对外合作”主题的情感波动较大，这可能与 5 月份峰会后中国和其他国家频繁签署协议有关，截至 7 月份总共与 61 个国家签署了 2 431 份合作协议，新签合同额 714.2 亿美元。7 月份起直到年底积极的情感值占比高于其他两项，对外方合作方面的消极情感值也有所上升。“‘一带一路’与经济效应”主题的积极情感值占比在大部分时间段高于其他两项。据商务部网站的介绍，2017 年中国企业共对“一带一路”沿线的 59 个国家非金融类直接投资 143.6 亿美元[33]，英文推文对“一带一路”带来的经济效应逐渐看好。“‘一带一路’与对外政策”主题在 5 月份高峰论坛后积极情感逐步上升，高峰论坛加强了政策沟通和战略对接，签署了多个双边、多边合作文件及企业合作项目[34]，得到了积极的反响。

4　总结

2017 年是“一带一路”倡议取得突破性进展的一年，“一带一路”国际合作高峰论坛召开，蒙内铁路正式通车、亚马尔液化天然气项目首条生产线投产等很多项目逐步落地，新的合作协议不断签署，“一带一路”一词成为全世界的热词，新闻报道更多的是从官方的角度呈现事实，而国内外对其反响难以明确，世界上众多用户在 Twitter 上对“一带一路”展开热议，从中更能体现用户的关注点和情感倾向。基于 LDA 的 Twitter 中英文文本分析发现，2017 年中文推文热议有六大主题，分别为：“一带一路”与朝鲜半岛和平问题、“一带一路”与经济问题、“一带一路”与高峰论坛之高层访问、“一带一路”与高峰论坛之合作与项目、“一带一路”与外交战略、“一带一路”与国内热点。英文推文热议有四大主题，分别为：“一带一路”与 CPEC、“一带一路”与对外合作、“一带一路”与经济效应、“一带一路”与对外政策。对比来看，中文的推文用户关注的问题较微观，对“一带一路”高峰论坛热议很高，更凸显“合作”的态度，以“合作”为出发点看待“一带一路”，看重“合作”的过程。英文推文关注的主题较宏观，对整体的趋势和发展讨论更多，更多从“一带一路”倡议的经济效应和发展情况进行评价。间接地反映出中文推文将“一带一路”倡议定位为“区域合作发展”项目，而英文推文把“一带一路”倡议视为“经济合作”项目。

从中文推文主题的情感演化模式看，除朝鲜半岛

和平问题外,其他主题的积极和中立情感均占主流,消极情感占比最少,且中立情感自 5 月份高峰论坛后均表现出下降趋势,用户的情感倾向更明确。英文推文的积极、中立的情感值也占据了主要位置,除"一带一路"与外交政策外,其他主题情感波动较小,波动点主要出现在 10 月份附近,可见十九大等 10 月份重要的节点事件对情感影响较大。本文通过主题挖掘和情感分析方法,尝试呈现国际社交媒体对"一带一路"倡议的关注内容和情感倾向。在未来还需要拓展到西班牙文、法文等多语种的推文,更全面地呈现国际上对"一带一路"倡议的响应态度和关注重点。

参考文献:

[1] 黄炎秋. 建构主义国际关系视域下"一带一路"对非洲传播策略研究[D]. 武汉:华中师范大学, 2017.
[2] 朱桂生,黄建滨. 美国主流媒体视野中的中国"一带一路"战略——基于《华盛顿邮报》相关报道的批评性话语分析[J]. 新闻界,2016(17):58-64.
[3] 龚言浩,甄峰,席广亮. "一带一路"倡议关注与响应的空间格局——基于新浪微博数据的分析[J]. 地域研究与开发,2018,37(2):29-35.
[4] 贾爽. "一带一路":Twitter 网络舆情分析与对策建议[D]. 南京:南京大学,2016.
[5] HOFMANN T. Probabilistic latent semantic indexing[J]. Sigir forum,2017, 51(2): 211-218.
[6] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(1): 993-1022.
[7] 陈晓美,高铖,关心惠. 网络舆情观点提取的 LDA 主题模型方法[J]. 图书情报工作,2015, 59(21):21-26.
[8] MICHELSON M, MACSKASSY S A. Discovering users' topics of interest on twitter: a first look[C]// Workshop on analytics for noisy unstructured text data. New York:ACM, 2010.
[9] YOOSUN H, HONGJIN S. Opinion leadership on twitter and twitter use - motivations and patterns of twitter use and case study of opinion leaders on twitter [J]. Korean journal of broadcasting and telecommunication studies, 2010, 24(6): 365-404.
[10] HU Y, JOHN A, SELIGMANN D D. Event analytics via social media[C]//ACM workshop on social and behavioural networked media access. New York:ACM, 2011:39-44.
[11] MEI Q, ZHAI C X. A mixture model for contextual text mining [C]//Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM, 2006: 649-655.
[12] MOGHADDAM S, ESTER M. ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews[C]//Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2011: 665-674.
[13] BLEI D M, LAFFERTY J D. Dynamic topic models[C]//Proceedings of the 23rd international conference on machine learning. New York:ACM, 2006: 113-120.
[14] ALSUMAIT L, BARBARÁ D, DOMENICONI C. On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking[C]//Eighth IEEE international conference on Data mining. Washington, DC: IEEE, 2008: 3-12.
[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in neural information processing systems. New York: Curran Associates, Inc. , 2013: 3111- 3119.
[16] GO A, BHAYANI R, HUANG L. Twitter sentiment classification using distant supervision[R]. Cs224n project report. Palo Alto: Stanford University, 2009.
[17] JANSEN B J, ZHANG M, SOBEL K, et al. Twitter power: tweets as electronic word of mouth[J]. Journal of the American Society for Information Science & Technology, 2009, 60(11):2169-2188.
[18] TUMASJAN A, SPRENGER T O, SANDNER P G, et al. Predicting elections with twitter: what 140 characters reveal about political sentiment[C]//International conference on weblogs and social media, Icwsm 2010. Washington, DC: DBLP, 2010.
[19] BOLLEN J, PEPE A, MAO H. Modeling public mood and emotion: twitter sentiment and socio-economic phenomena[J]. Computer science, 2009, 44(12):2365-2370.
[20] SHERIDAN D P, DECKER H K, KLOUMANN I M, et al. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter[J]. PlOS ONE, 2011, 6(12): e26752.
[21] HU M, LIU B. Mining and summarizing customer reviews[C]// Tenth ACM SIGKDD international conference on knowledge discovery and data mining. Seattle:ACM, 2004:168-177.
[22] ZHANG L, GHOSH R, DEKHIL M, et al. Combining lexicon-based and learning-based methods for Twitter sentiment analysis [EB/OL]. [2018-04-03]. https://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf.
[23] 知网发布情感分析用词语集[EB/OL].[2018-03-08]. http://www.keenage.com/html/c_bulletin_2007.htm.
[24] FELDMAN R. Techniques and applications for sentiment analysis [M]. New York:ACM, 2013.
[25] VOLKOVA S, WILSON T, YAROWSKY D. Exploring sentiment in social media: bootstrapping subjectivity clues from multilingual twitter streams[C]// Proceedings of ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics. Sofia: ACL,2013:505-510.
[26] VOLKOVA S, WILSON T, YAROWSKY D. Exploring demographic language variations to improve multilingual sentiment analysis in social media[C]//Proceedings of conference on empirical

methods in natural language processing. Seattle:ACL,2013: 1815－1827.

[27] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on empirical methods in natural language processing-Volume 10. Philadelphia:ACM,2002: 79－86.

[28] 张志华. 基于深度学习的情感词向量及文本情感分析的研究[D].武汉:华东师范大学,2016.

[29] 范云满，马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报，2014，33(7):698－711.

[30] 贺亮，李芳. 基于话题模型的科技文献话题发现和趋势分析[J]. 中文信息学报，2012，26(2):109－116.

[31] 杨希. 基于情感词典与规则结合的微博情感分析模型研究[D].合肥:安徽大学,2014.

[32] 法媒称中巴经济走廊重振巴基斯坦[EB/OL].[2018－04－03]. http://news.163.com/17/0805/00/CR1LUS9A00018AOQ_all.html.

[33] 2017 年我对"一带一路"沿线国家投资合作情况[EB/OL].[2018－04－03]. http://fec.mofcom.gov.cn/article/fwydyl/tjsj/201801/20180102699450.shtml.

[34] "一带一路"国际合作高峰论坛"政策沟通"平行主题会议签署 32 个合作协议[EB/OL].[2018－04－03]. http://www.xinhuanet.com/2017-05/14/c_1120970716.htm.

作者贡献说明:

赵常煜:确定选题,提出论文研究框架,撰写论文;

吴亚平:论文修改、框架调整;

王继民:提出研究思路,修订论文。

Twitter Text Topic Mining and Sentiment Analysis Under the Belt and Road Initiative

Zhao Changyu[1]　Wu Yaping[2]　Wang Jimin[1]

[1] Department of Information Management, Peking University, Beijing 100871

[2] Peking University Library, Beijing 100871

Abstract: [**Purpose/significance**] The Belt and Road Initiative has attracted widespread attention around the world, and users in many countries have expressed their opinions, comments and discussed with each other on twitter, the most representative social media. The discussion topic and emotional tendency of "the Belt And Road" in the world extracted from the tweets will be helpful for the government to optimize their propaganda strategies and increase the exposure and attention of the Belt and Road Initiative. [**Method/process**] This paper collected more than 60 000 tweets related to the Belt and Road Initiative in 2017, and respectively carried out data preprocessing, data description, topic mining, and sentiment analysis in Chinese and English, and realized cross-analysis of topics and emotions to draw conclusions. [**Result/conclusion**] The tweet theme in 2017 is mainly around the "Belt and Road Forum for International Cooperation". Chinese tweets pay more attention to the planning and implementation of the forum, as well as security issues, visits by the leadership, etc. The emotional value fluctuates greatly, especially the negative emotions on security issues. English tweets are more concerned with the facts of holding the summit forum and the economic effects brought by the forum. The emotional fluctuations are small, and the emotional value of the economic aspect is that the positive proportion is significantly higher than the negative and neutral emotional values.

Keywords: the Belt and Road Initiative　twitter　topic mining　sentiment analysis